Thirukkural - A Text-to-Speech Synthesis System

G. L. Jayavardhana Rama, A. G. Ramakrishnan, M Vijay Venkatesh, R. Murali Shankar

Department of Electrical Engg, Indian Institute of Science, Bangalore 560012, INDIA

1. Introduction

In this paper, we propose a novel method for Text-To-Speech (TTS) conversion in Tamil language. It involves two phases, namely, the offline phase and the online phase. Offline phase includes pre-processing, segmentation and pitch marking. Online phase includes text analysis and synthesis. The paper is organized as follows: Section 2 reviews the different methods for synthesizing speech. Section 3 explains the offline processes and the online phase is described in section 4. Section 5 discusses the implementation and section 6 describes the potential applications. Section 7 gives the conclusion.

2. Speech Synthesis

The techniques employed for synthesizing speech from text may be broadly classified into three categories:

I. Formant-based
II. Parameter-based
III. Concatenation-based

In the formant-based approach [1], we can synthesize a signal by passing the global periodic waveform through a filter with the formant frequencies of the vocal tract. It makes use of the rules for modifying the pitch, formant frequencies and other parameters. However, the technique fails to produce good quality, natural sounding speech, as appropriate rules cannot be derived for unlimited speech. As the model uses a number of resonators, it is computationally expensive.

On the other hand, in parameter-based synthesizers [1], the waveforms are modelled using Linear Prediction (LP) coefficients. These LP coefficients fail to model nasals perfectly. The linear prediction model is an all-pole model which models vowels exceptionally well, but fails to model the nasals and silence (stops).

In concatenation-based speech synthesis, natural speech is concatenated to give the resulting speech output. This is more natural but the database size is fairly huge. Concatenation method can be of three different types:

(a) Limited domain waveform concatenation

For a given limited domain, this approach can generate very high quality speech with only a small number of recorded segments. Such an approach, used in most interactive voice response systems, cannot synthesize arbitrary text. Many concept-to-speech systems use this approach.

(b) Concatenation without waveform modification

Unlike the previous approach, these systems can synthesize speech from arbitrary text. They can achieve good quality on a large set of sentences, but the quality can be mediocre for other sentences where poor concatenations take place.

(c) Concatenation with waveform modification

These systems have more flexibility in selecting the speech segments to concatenate because the waveforms can be modified to allow for a better prosody match. This means that the number of sentences with mediocre quality is lower than the case where no prosody modification is allowed. On the other hand, replacing natural with synthetic prosody can hurt the overall quality. In addition, the prosody modification process also degrades the overall quality. Figure1 shows the block diagram of our TTS system using concatenation principles.

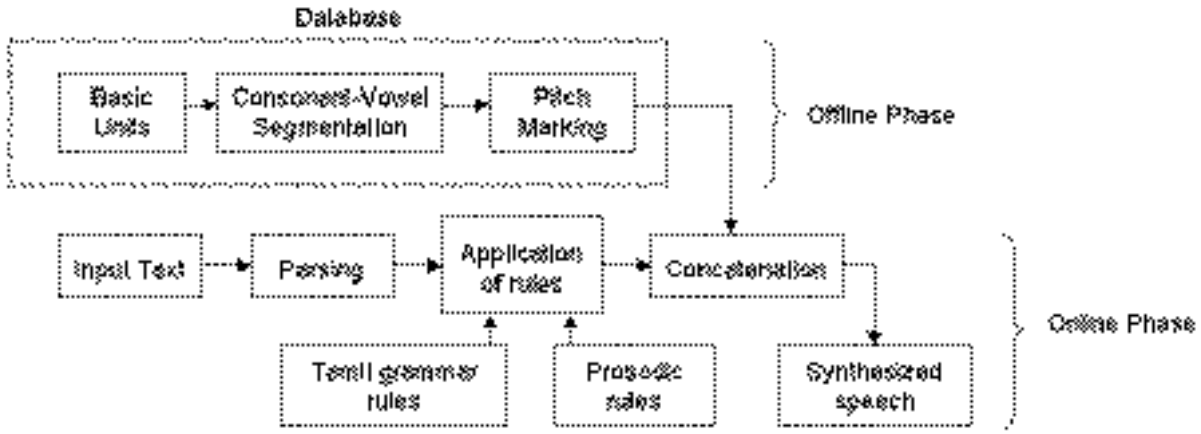

Figure 1. Block diagram of speech synthesis system using concatenation

In our work, we have employed the second of the above methods, that is, concatenation without waveform modification. Analysing the advantages and disadvantages of such a method, we proceed to propose a method to improve the quality of the synthesized speech.

3. Offline process

The offline processes of the system include (1) Choosing the basic units (2) Building the database (3) Detailed study of prosody in natural speech (4) Consonant - vowel segmentation (5) Pitch marking.

3.1 Deciding the basic units

Basic unit of speech is a phoneme. Other units that can be used for synthesis are diphones, triphones, demi-syllables, syllables, words, phrases and sentences. In terms of quality, sentence is the best and phoneme is the worst. However, the size of the database also is an important factor to

be considered. Practically infinite units must be stored if the basic units are sentences. The issues in choosing the basic units for synthesis are:

The units should lead to low concatenation distortion. A simple way of minimizing this distortion is to have fewer concatenations and thus use long units such as words, phrases or even sentences. However, as described in the previous paragraph, the size of the database size can become unwieldy.

The units should lead to low prosodic distortion. Whereas it is all right to have units with prosody slightly different from the desired target, replacing a unit having a rising pitch with another having a falling pitch will result in an unnatural sentence.

The units should be of a general nature, if unrestricted text-to-speech is required. For example, if we choose words or phrases as our units, we cannot synthesize arbitrary speech from text, because it is almost guaranteed that the text will contain words not in our inventory. As an example, the use of arbitrarily long units results in low concatenation distortion but it has been shown that over 180,000 such units would be needed to cover 75% of a random corpus. The longer the speech segments are, the more of them we need, to be able to synthesize speech from arbitrary text. This generalization property is not needed if closed-domain synthesis is desired.

Considering the above issues, syllables have been used as basic units. This may contain phonemes, diphones or triphones. The different instances of the unit are V, CV, VC, VCV, VCCV and VCCCV, where V stands for a vowel and C stands for a consonant.

Building the database

The database was collected from a native Tamil speaker over a span of several months. Recording took place in a noise free room using Shure SM 58 microphone, whose frequency response is 50-15,000 Hz [2], connected to a PIII - 500 MHz PC. Spoken units were recorded at a sampling rate of 8 KHz.

3.2 Observation of prosody in natural speech

Prosody is a complex weave of physical, phonetic effects that is being employed for expression. Prosody consists of systematic perception and recovery of the speaker‍ s intentions based on pauses, pitch, duration and loudness. Pauses are used to indicate phrases and to indicate end of sentence or paragraph. It has been observed that the silence in speech increases as we go from comma to end of sentence to end of paragraph. Pitch is the most expressive part of a speech signal. We try to express our emotion through pitch variation. Constant pitch signal sounds very unnatural. In the Tamil language, sudden variation in pitch does not occur in a vowel as it happens in Japanese. Duration is the second important factor that affects the naturalness of the synthesized speech. Same vowel appearing in different positions in a word or a sentence has different durations. For example, consider the sentence " /naan aaru manikku varalaamaa/ ". In this sentence, vowel /aa/ appears at different positions as tabulated below:

Word	Position	Duration (Sec)
Naan	Middle	0.15
aaru	Initial	0.16
Varalaamaa	Middle	0.15
Varalaamaa	Final	0.18

Table 1:Duration of vowel /aa/ when it appears at different positions

Duration analysis as shown above is performed on a set of samples recorded from a native Tamil speaker. The duration information is tabulated and is stored as a look up table for future reference. Although loudness is not as important a phenomenon as pitch, it may introduce artifacts. The artifact is in the form of an echo. This is caused due to the amplitude envelope mismatch.

3.3 Consonant Vowel Segmentation

It is observed that any change in the consonant part of a signal results in change of perception of the unit. Consonants must be kept intact. To this end, consonant and the vowel regions of the units must be segmented. In terms of morphological structure, consonant can be classified into co-articulated and non co-articulated signal.

Non co-articulated consonant can be segmented easily using difference of energy in consecutive blocks of the signal. The given speech unit is divided into frames of 10 msec duration each. Energy of each frame is calculated and the first difference of the energy contour gives two distinct peaks, one on the positive side and the other, on the negative side.

Co-articulated consonant segmentation is more challenging than the other. Energy contour is almost flat at the transition from vowel to consonant. Preliminary results based on spectral analysis are available. For completion of the project, we resorted to manual segmentation.

3.4 Pitch Marking

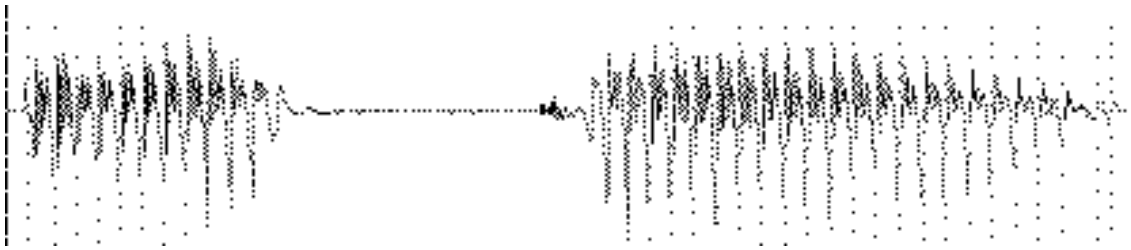

Figure 2. Segmented and pitch marked non-co-articulated VCV /aka/

Pitch marking is essential as the waveforms are concatenated at the pitch marks. The method employed for estimation of pitch is auto-correlation. The length of each period is calculated by finding the maximum autocorrelation in a given segment. Segment length is taken to be 10 msec. After getting the peaks, nearest zero crossing to the left of the peak gives the pitch mark. Unbiased autocorrelation is employed to get distinct peaks at the pitch frequency. The results of pitch marking and segmentation of non-co-articulated and co-articulated consonants are as shown in Figs. 2 and 3, respectively.

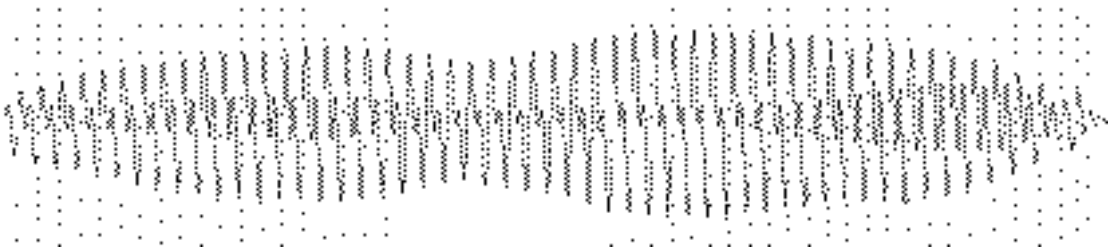

Figure 3. Segmented and pitch marked co-articulated VCV /iyi/

4. Online Process

This involves two important phases: (a) Text analysis (b) Synthesis. Text analysis phase involves parsing the input text into a sequence of basic units of speech and application of Tamil rules. Synthesis phase involves the concatenation of the waveforms of these units in the correct sequence and synthesis after application of the prosodic rules. Prosodic rules involve duration adjustment followed by amplitude modification. Duration is adjusted as per the durations in the look up table as obtained in section 3.3. The nearby pitch mark is used as concatenation point. Pitch is generally constant at the concatenation points as the recording person is asked to record at constant pitch. A small variation in pitch at the concatenation point does not affect the quality of speech. At the point of concatenation there may be a mismatch in the amplitude of the two waves. This will sound like an echo after concatenation. This is normalised by fixing a threshold and matching the amplitude of the vowels.

5. Implementation

The system is designed to work on Windows 95, Windows 98 and Windows NT. It is designed using C++ and graphic user interface (GUI) is provided using Visual C++. The output is stored in wave file format. It supports TAB keyboard as recommended by Tamilnet99 standards. Phonetic Tamil keyboard has also been implemented. Figure 4 shows the appearance of the software.

6. Applications

The applications of speech synthesis in Tamil are listed below:

- Natural language interface for computers
- Self-learning multimedia education packages in Tamil
- Automatic telephone-based enquiry systems in all Government organisations
- Audio on-line help in all Indian language based IT software
- Computer based language teaching
- Automatic document reading machines in Indian languages for the blind
- To synthesize different types of voices in animation movies
- Digital Personal Assistant (English/Indian language to Indian Language).

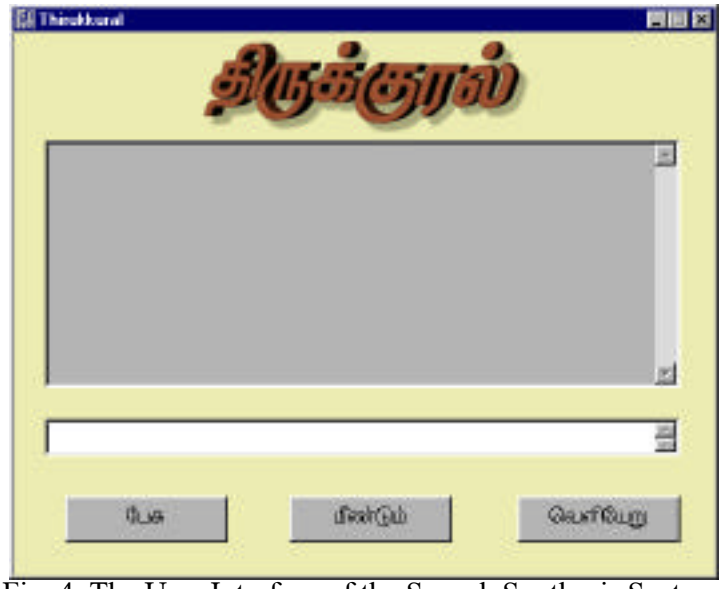

Fig. 4. The User Interface of the Speech Synthesis System.

7. Conclusion

Thirukkural synthesizes intelligible Tamil speech. It has a male voice and can read input text, which is in TAB format. Although the techniques used are quite primitive, quality of speech is good. Efforts are on to reduce the size of the database, which is currently high. Attempts are being made to make it natural, add emotions, making it net enabled and also to provide good synthesis for alien words (like /fa/ in 'father').

Acknowledgement

We thank the Government of Tamilnadu for funding part of this project through ELCOT under the Tamil Software Development Fund.

References

[1] Douglas O'Shaughnessy (2000), Speech Communication - Human and Machine
[2] Shure SM58 Technical Manual